OKU Fen Bilimleri Enstitüsü Dergisi
7(1): 414-423, 2024

OKU Journal of The Institute of Science and Technology, 7(1): 414-423, 2024

Osmaniye Korkut Ata Üniversitesi
Fen Bilimleri Enstitüsü
Dergisi

Osmaniye Korkut Ata University
Journal of The Institute of Science
and Technology

İngilizce Metinlerin Sınıflandırması İçin Makine Öğrenimi Kullanımı (A mini-review)

Somayyeh SHABESTANI[1*], Merve GEÇİKLİ [2]

[1]Department of Foreign Languages Teaching English Language Teaching, Faculty of Education, Ataturk University, Erzurum 25240, Türkiye
[2]Department of Foreign Languages Teaching English Language Teaching, Faculty of Education, Ataturk University, Erzurum 25240, Türkiye

[1]https://orcid.org/0000-0001-6773-869X
[2]https://orcid.org/0000-0002-8619-5026
*Sorumlu yazar: somayyeh.shabestani20@ogr.atauni.edu.tr

Derleme

Makale Tarihçesi:
Geliş tarihi: 03.03.2023
Kabul tarihi:06.06.2023
Online Yayınlanma: 22.01.2024

Anahtar Kelimeler:
Sınıflandırma
İngilizce metinler
Makine öğrenimi

ÖZ

Metinden verilerin elde edilmesi ve bu verilerden bilgi elde etmek için metin sınıflandırma sisteminin kullanılması, metnin içeriğinin anlamlandırmasında okuyucunun yorum gücünü geliştirir. Bu bağlamda, metinlerin zorluk düzeylerine göre sınıflandırılmasında, teknolojik gelişmeler ışığında, önemli gelişmeler yaşanmış ve yeni sistemler ortaya çıkmıştır. İngiliz dilbilimi temelli çalışmalarda da temel odak noktası ve en yaygın problemlerden biri, metinleri zorluk düzeylerine göre sınıflandırmaktır. Buradan hareketle, bu makalenin temel amacı günümüzde İngilizce metin sınıflandırmasında en çok kullanılan yeni sistemlerden birini tanıtmaktır. Bu bağlamda, çalışmanın temel odağı İngilizce metinlerin zorluk/okunabilirlik açısından sınıflandırılmasında kullanılan makina öğrenim algoritmaları ile ilgili bir mini inceleme yapmaktır. Ayrıca, bu algoritmaların güçlü ve zayıf yönleri de ele alınacaktır.

Machine Learning Use For English Texts' Classification (A mini-review)

Review Article

Article History:
Received: 03.03.2023
Accepted: 06.06.2023
Published online: 22.01.2024

Keywords:
Classification
English texts
Machine learning

ABSTRACT

Using classification to retrieve information and extract data from text increases the reader's understanding of the content. In this regard, thanks to technological advances, text classification systems have been updated and new systems have emerged. The main focus and one of the common problems observed in English Linguistics studies is to classify texts to the readability level. Thus, the main aim of this study is to shed light on one of the new systems commonly used today in textual classification of English texts. In this regard, the main focus of the paper is to provide a mini review of the sort of machine learning algorithms used in classifying English text regarding difficulty/readability level. Besides, weak and strong sides of these algorithms will also be mentioned in detail.

To Cite: Shabestani S., Geçikli M. Machine Learning Use For English Texts' Classification (A mini-review). Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2024; 7(1): 414-423.

1. Introduction

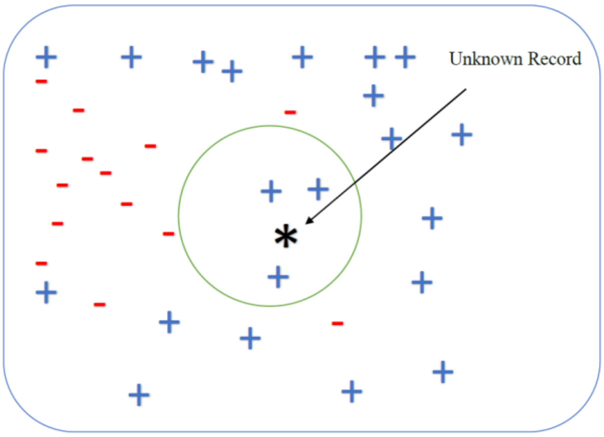
Text classification is also known as text tagging or text categorization and is considered an "interdisciplinary" issue expertise in the following fields: library science, computer science, and information science (Kavitha and Prabhavathy, 2021). Text classification aims to systematically classify or categorize text data into specific and predetermined groups (Wu et al., 2014; Wahdan et al., 2020). Text classification methods can analyze text data automatically by using natural language processing algorithms and create predefined labels, classes, or classifications based on the content of the data. The concept of text classification was first expressed in the early sixties (Xia and Du, 2011), and focused on indexing scientific journals using vocabulary. It is important to classify texts gathered from a broad range of data sources (such as social networks, websites, and published online research papers) in order to better understand them. (Altınel and Ganiz, 2018; Liu et al., 2018). Manual classification is difficult and time-consuming but inaccurate (Wang et al., 2018; Yu et al., 2019). Advanced methods have been developed to solve these problems and make text classification more reliable. Using intelligent algorithms, automatic text classification (such as machine learning and natural language processing) can be performed more accurately, quickly, and cost-effectively. (Hirway et al., 2022). Over the past 20 years, digital texts have been rapidly created, making it difficult to find contextual documents that require a deeper understanding of machine learning techniques. (Kowsari et al., 2019). Text classification methods include searching for similar documents, classifying them according to their topics, and creating new documents. Although many educational applications of text mining have been published recently, we have not found any paper examining them in the English linguistics field. Accordingly, this work presents an overview of the current status of the classified text into different categories in the English linguistics field.

2. Text classification

To improve text classification accuracy, in a study by Ghareb and others (2016) an updated feature was presented selection method based on a genetic algorithm. In order to deal with high dimensions of the space feature and improve the performance of classification simultaneously, the following method combines the selection of features filter techniques with the improvement of the genetic algorithm. The method of extracting keywords from texts and classifying them in a classification of texts is the focus of Onan et al. (2016) study. Term identification is crucial to analytics, natural language processing, and information retrieval. By analyzing keywords, we are able to access compressed text documents. A compact representation of documents can include automated indexing, intelligent summarization, auto classification, categorization, and clarifying techniques. A study by Shafiabady et al. (2016) found that unsupervised clustering using support vector machines can be used to classify text for the training phase of text classification algorithms, many tagging and management tasks must be completed by humans. The use of expert systems is essential in all cases except when dealing with textual documents, since manually classifying and tagging large numbers of documents can be both quite time- consuming and intellectually exhausting. Moreover, some new fields may not know how to organize and summarize different classes, so an unsupervised training scheme is required for automatic clustering data. Elghazel et al. (2016) have investigated the classification of texts using multiple labeling and have concluded that a document can belong to several classifications at the same time (e.g., viruses, The health field, athletics, and the Games of the Olympics), and textual analysis Offers many possibilities to develop new multi-label learning methods, especially for text data. Uysal (2016) developed an improved feature selection scheme for text classification, and filter-based feature selection methods are preferred for text classification due to their effectiveness. The combined feature selection scheme assigns scores to each feature based on their discriminating power and then sorts them descendingly. A group of elements is then improved by adding N- to them, where N is an empirical number. This paper presents the final step in the joint feature selection scheme to refine the combined set of features by introducing an improved general feature selection scheme (IGFSS). Using fuzzy evolving grammar for classifying crime texts, Scharf and Martin (2015) define analysis as the activity of identifying useful information in natural language texts. The solution to this general problem is machine learning-based methods. This is a crucial criterion for text classification. A recent article by Zhang et al. (2016) explored sentiment prediction using particle swarm optimization and feature selection search in an attempt to supply a human-understandable representation. Using particle clustering-based factor extraction in medical analysis and forecasting emotions into textual documents as a machine learning technique to gain insights from unstructured texts has gained new popularity in the healthcare industry in recent years. A new learning class built on a resource allocation-based learning network (SLRAN) for text classification is presented by Song et al. (2015). This study examined the utilization of learning-base network resources for automatic text classification. Based on the progress of learning, SLRAN is divided into initial and enhanced learning sections. As Thomas and Resmipriya (2016) explored the effect of text classification with a clustering scheme, this article examines how to achieve better performance using a classification method based on similarity and efficiency criteria. An additional step to classifying and defining text collection components is semi-supervised clustering. Labeled texts are classified to identify text clusters and unlabeled texts that match centers. The categories of each text cluster are read with the labels of texts in it. The text clusters are then used to generate the model for classification and the next steps for text classification. The rule-based approach organizes texts into organized groups by using linguistic rules to analyze contextual sentiment (polarity) (Neviarouskaya et al., 2011; Shahi and Sitaula, 2021). In spite of this, there is little information about text classification. This study presents primary methods for the classification of the text. Several popular text classifiers are discussed next, Nearest Neighbor (NN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Neural Networks are some of them.

3. Nearest Neighbor method

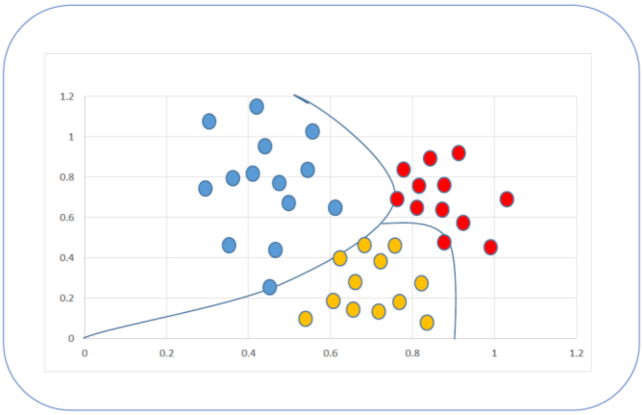
In the nearest neighbor method, k records are selected from the set of training records closest to the test record. The test record category is determined by the category's superiority or label corresponding to the test record. This method selects the category that has the most records in the selected neighborhood for that category. Therefore, the category that is most frequently observed among all the categories and the nearest neighbors is considered a new record.



Its high computational cost is a major drawback of the NN method, even though it is widely used in many real-world applications. A NN method uses lazy learning and finds the k nearest neighbors every time an object is given based on its k nearest neighbors.

4. Naïve Bayes Approach

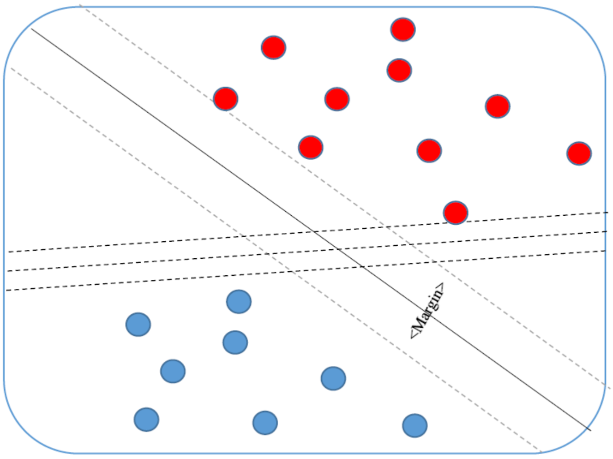
Naive Bayes is another classification method used in text classification. It is a simple probabilistic classifier with a learning step that estimates the probability of occurrence and is used to classify the newly created sample (Harahap et al., 2018). Ordinarily, the supposition accepted by NB classifiers is that particular feature's value and any other feature is independent. Generally, each word within a document has an independent probability of being discovered of the presence of other words within it (Deng et al., 2019).



A study by Kim et al. (2006) found that NB classifiers have two drawbacks: they have a rough estimation of parameters, as well as a bias against classes with a small number of training documents.

5. Method based on Support Vector Machines (SVM)

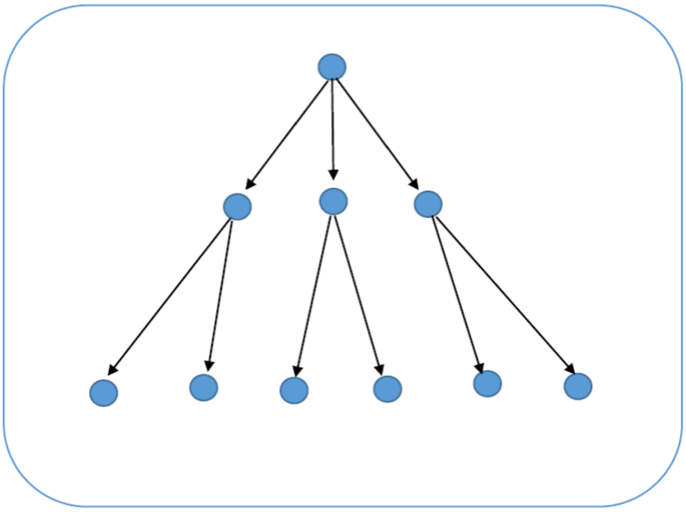
Recently, support vector machine algorithms have received extensive attention for text classification (Rustam and Yaurita, 2018; Park et al., 2020). Vapnik et al. first proposed this algorithm in 1992 (Boser et al., 1992), establishing the statistical theory of learning (Wan et al., 2012). General properties of support vector machines are 1) Maximizing generalization in a classifier, 2) Finding the global optimum for the cost function, 3) An automatic algorithm for determining the optimal classifier structure and topology, and 4) Using the Hilbert space and the inner product to model nonlinear discriminant functions.



A SVM approach in the learning phase aims to maximize the minimum distance between categories by choosing the decision boundary. Noise conditions in practicing this kind of choice can be tolerated. This border selection method is based on support vectors (Cervantes et al., 2020). This type of border selection is done based on points called support vectors.

6. Decision Tree methods

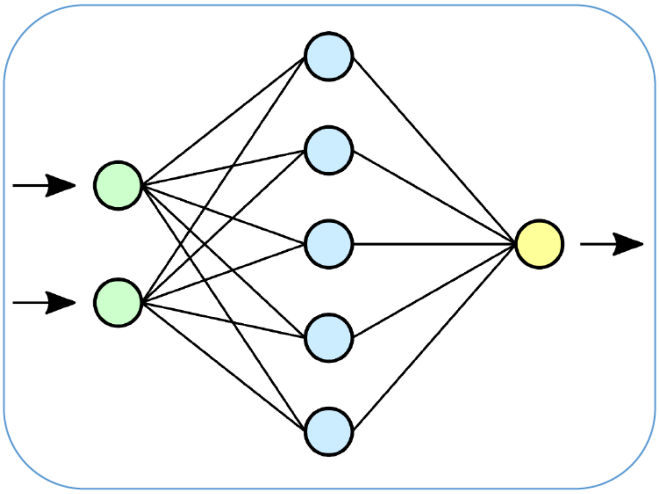
Decision trees are built using a divide and conquer strategy. As an example, we have a training set of tagged documents to split. The word ti should be chosen as a criterion for partitioning the training set. A term is selected, then M is divided into two subsets based on the term selected. The subset Mi+ contains documents that contain M, the subset Mi- contains documents that do not contain M. Repeated processes for Mi+ and Mi- are followed until all the documents within a subset have the same class (Brunello et al., 2018; Neelakandan and Paulraj, 2020).



A decision tree is a standard tool in data mining (Nourani et al., 2019). These algorithms are fast and scalable in terms of variables and training sets. However, their reliance on a small number of terms makes them challenging to use for text mining. A boosting method, for instance, uses multiple complementary decision trees to reduce error better than a decision tree method (Thangaraj et al., 2018; Cai et al., 2019).

7. Neural Networks Methods

In response to these difficulties, experts have developed more advanced mechanisms for text classification based on traditional formulas. Also with the advancement of information and communication technologies, machine learning algorithms, known as data-based intelligent software systems (As published by Li et al. in 2018; Basiri and colleagues., 2020), have emerged. Classification of text by machine learning, past observations are used to make classifications instead of manually crafting rules. The machine learning algorithm is capable of learning the relationships between text pieces by using pre-labeled examples as training data. It can also predict which output (tags) is expected for which input (text). Texts are classified according to a predetermined category or classification called a tag. Feature extraction is the initial step in training a machine learning classifier, which converts text into numerical vectors. Among the most commonly used feature extraction methods is the bag of words approach, in which a vector represents a word's frequency within a dictionary.



8. Conclusion

A valid and reliable definition of sentence complexity has been developed in recent years due to further developments in the index-based methods driven by data (machine-learning-based). In addition, this definition takes into account complex features such as the frequency of words, lexical meaning, morphology of the text, and parsing depth. The sentence complexity is calculated based on SVM. The system categorizes texts based on lexical, syntactic, and morpheme features. As natural language processing techniques and machine learning algorithms gain popularity, researchers can refine model algorithms to measure text readability with more flexibility.

Disputes of Interest

There is no conflict of interest between the authors. The manuscript was submitted without a conflict of interest, and all authors approved its publication.

Authorship Contributions

S.S. initiated the project idea. S.S. and M.G. contributed to the idea, design, and execution of the study. S.S. drafted and finalized the manuscript.

References

Altınel B., Ganiz MC. Semantic text classification: A survey of past and recent advances. Information Processing and Management 2018; 54(6): 1129-1153.

Basiri ME., Abdar M., Cifci MA., Nemati S., Acharya UR. A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. Knowledge-Based Systems 2020; 198: 105949.

Boser BE., Guyon IM., Vapnik VN. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory 1992; 144-152.

Brunello A., Marzano E., Montanari A., Sciavicco G. J48S: A sequence classification approach to text analysis based on decision trees. In International Conference on Information and Software Technologies 2018; 240-256, Springer, Cham.

Cai L., Gu J., Ma J., Jin Z. Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees. Energies 2019; 12(1): 159.

Cervantes J., Garcia-Lamont F., Rodríguez-Mazahua L., Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 2020; 408: 189-215.

Deng X., Li Y., Weng J., Zhang J. Feature selection for text classification: A review. Multimedia Tools and Applications 2019; 78(3): 3797-3816.

Elghazel H., Aussem A., Gharroudi O., Saadaoui W. Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. Expert Systems with Applications 2016; 57: 1-11.

Ghareb AS., Bakar AA., Hamdan AR. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications 2016; 49: 31-47.

Harahap F., Harahap AYN., Ekadiansyah E., Sari RN., Adawiyah R., Harahap CB. Implementation of Naïve Bayes classification method for predicting purchase. In 2018 6th International Conference on Cyber and IT Service Management (CITSM) 2018; (pp. 1-5). IEEE.

Hirway C., Fallon E., Conolly P., Flanagan K., Yadav D. Determining receipt validity from e-mail subject line using feature extraction and binary classifiers. International Journal of Simulation--Systems, Science and Technology 2022; 23(2).

Kavitha M., Prabhavathy P. A review on machine learning techniques for text classification. In 2021 4th International Conference on Computing and Communications Technologies (ICCCT) 2021; (pp. 605-610). IEEE.

Kim SB., Han KS., Rim HC., Myaeng SH. Some effective techniques for naive bayes text classification. IEEE Transactions on Knowledge and Data Engineering 2006; 18(11): 1457-1466.

Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., Brown D. Text classification algorithms: A survey. Information 2019; 10(4): 150.

Li C., Zhan G., Li Z. News text classification based on improved Bi-LSTM-CNN. In 2018 9th International conference on information technology in medicine and education (ITME) 2018; (pp. 890-893). IEEE.

Liu CZ., Sheng YX., Wei ZQ., Yang YQ. Research of text classification based on improved TF-IDF algorithm. In 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE) 2018; (pp. 218-222). IEEE.

Neelakandan S., Paulraj D. A gradient boosted decision tree-based sentiment classification of twitter data. International Journal of Wavelets, Multiresolution and Information Processing 2020; 18(04): 2050027.

Neviarouskaya A., Prendinger H., Ishizuka M. Affect analysis model: novel rule-based approach to affect sensing from text. Natural Language Engineering 2011; 17(1): 95-135.

Nourani V., Tajbakhsh A.D., Molajou A. Data mining based on wavelet and decision tree for rainfall-runoff simulation. Hydrology Research 2019; 50(1): 75-84.

Onan A., Korukoğlu S., Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications 2016; 57: 232-247.

Park K., Hong JS., Kim W. A methodology combining cosine similarity with classifier for text classification. Applied Artificial Intelligence 2020; 34(5): 396-411.

Pradhan A. Support vector machine-a survey. International Journal of Emerging Technology and Advanced Engineering 2012; 2(8): 82-85.

Rustam Z., Yaurita F. Insolvency prediction in insurance companies using support vector machines and fuzzy kernel c-means. In Journal of Physics: Conference Series 2018; 1028(1): 012118. IOP Publishing.

Shafiabady N., Lee LH., Rajkumar R., Kallimani VP., Akram NA., Isa D. Using unsupervised clustering approach to train the Support Vector Machine for text classification. Neurocomputing 2016; 211: 4-10.

Shahi TB., Sitaula C. Natural language processing for Nepali text: a review. Artificial Intelligence Review 2021; 1-29.

Song W., Chen P., Park SC. Application of a staged learning-based resource allocation network to automatic text categorization. Neurocomputing 2015; 149: 1125-1134.

Thangaraj M., Sivakami M. Text classification techniques: A literature review. Interdisciplinary Journal of Information, Knowledge, and Management 2018; 13: 117.

Thomas AM., Resmipriya MG. An efficient text classification scheme using clustering. Procedia Technology 2016; 24: 1220-1225.

Tran CK., Ngo TH., Nguyen CN., Nguyen LA. SVM-based face recognition through difference of Gaussians and local phase quantization. International Journal of Computer Theory and Engineering 2021; 13(1): 1-8.

Uysal AK. An improved global feature selection scheme for text classification. Expert systems with Applications 2016; 43: 82-92.

Wahdan KA., Hantoobi S., Salloum SA., Shaalan K. A systematic review of text classification research based ondeep learning models in Arabic language. International Journal of Electrical and Computer Engineering 2020; 10(6): 6629-6643.

Wan CH., Lee LH., Rajkumar R., Isa D. A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. Expert Systems with Applications 2012; 39(15): 11880-11888.

Wang WM., Li Z., Tian ZG., Wang JW., Cheng MN. Extracting and summarizing affective features and responses from online product descriptions and reviews: A Kansei text mining approach. Engineering Applications of Artificial Intelligence 2018; 73: 149-162.

Wu Q., Ye Y., Zhang H., Ng MK., Ho SS. ForesTexter: an efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems 2014; 67: 105-116.

Xia T., Du Y. Improve VSM text classification by title vector based document representation method. In 2011 6th International Conference on Computer Science & Education (ICCSE) 2011; (pp. 210-213). IEEE.

Yu M., Huang Q., Qin H., Scheele C., Yang C. Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and Irma as case studies. International Journal of Digital Earth 2019; 12(11): 1230-1247.

Zhang Y., Szabo C., Sheng QZ. Improving object and event monitoring on Twitter through lexical analysis and user profiling. In International Conference on Web Information Systems Engineering 2016; (pp. 19-34). Springer, Cham.